Replica Consistency Issues in Big Data Systems

NIST Big Data Public Working Group
IEEE Big Data Workshop
October 27, 2014

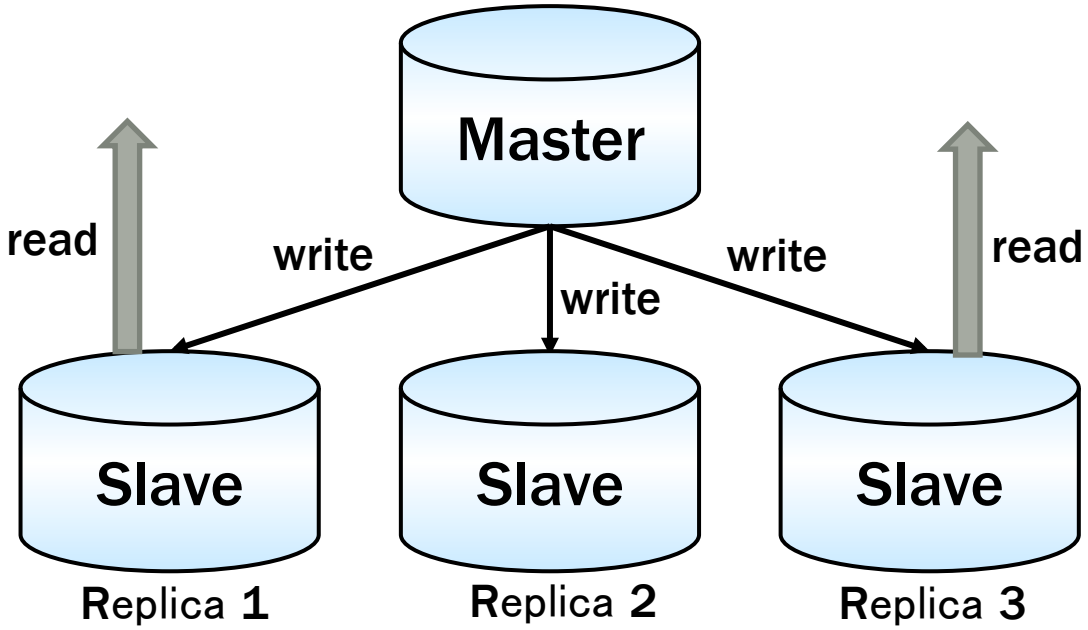
Jianmin Wang
School of Software, Tsinghua University
jimwang@tsinghua.edu.cn

Overview

- Objectives
- Approach
- Progress
- Next Steps

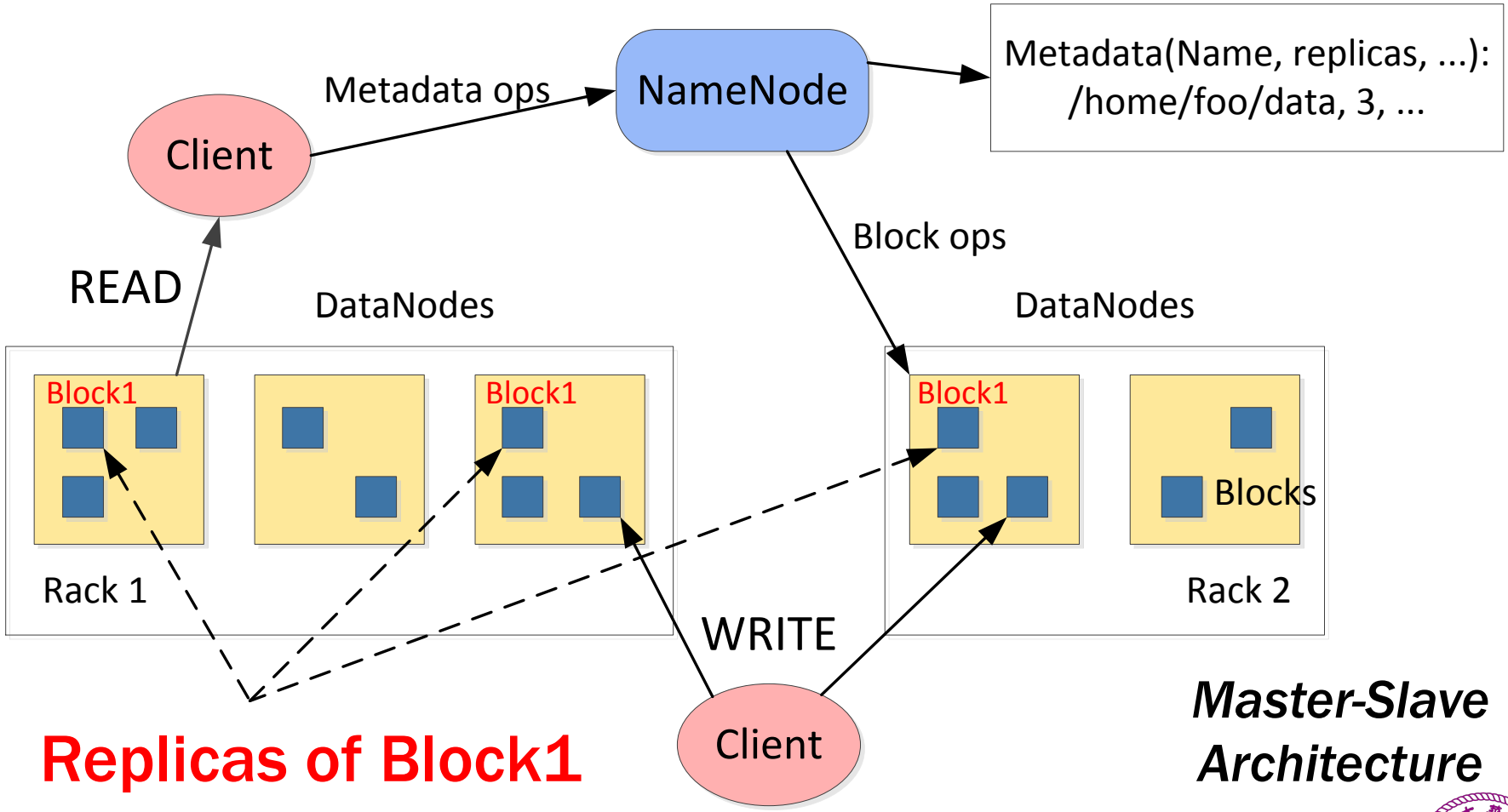
Data Replica

- Distributed data storage system is ubiquitous in big data applications
- **Replica mechanism** improves the system performance and reliability

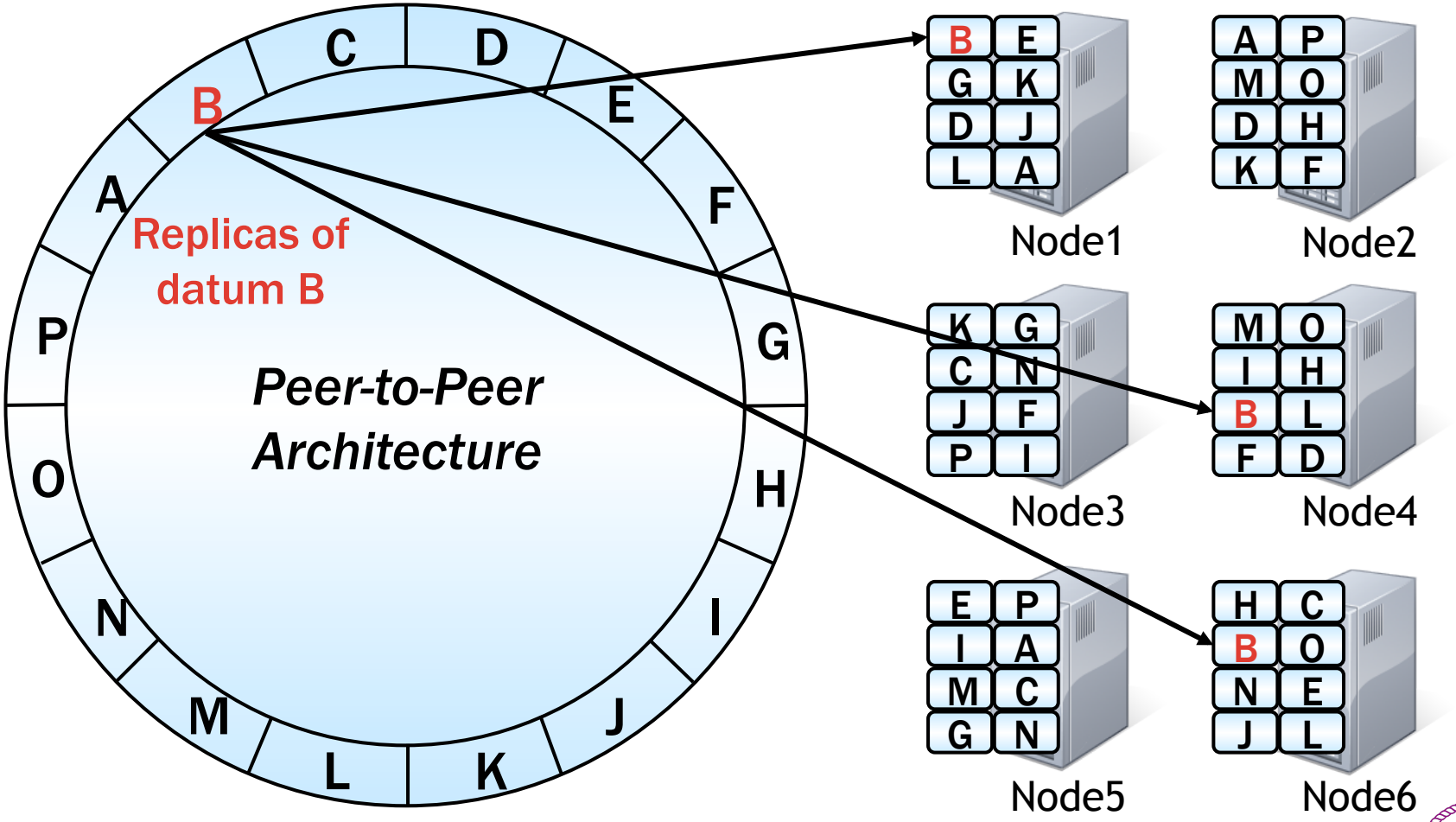


GB→TB→PB

Data Replica in Hadoop

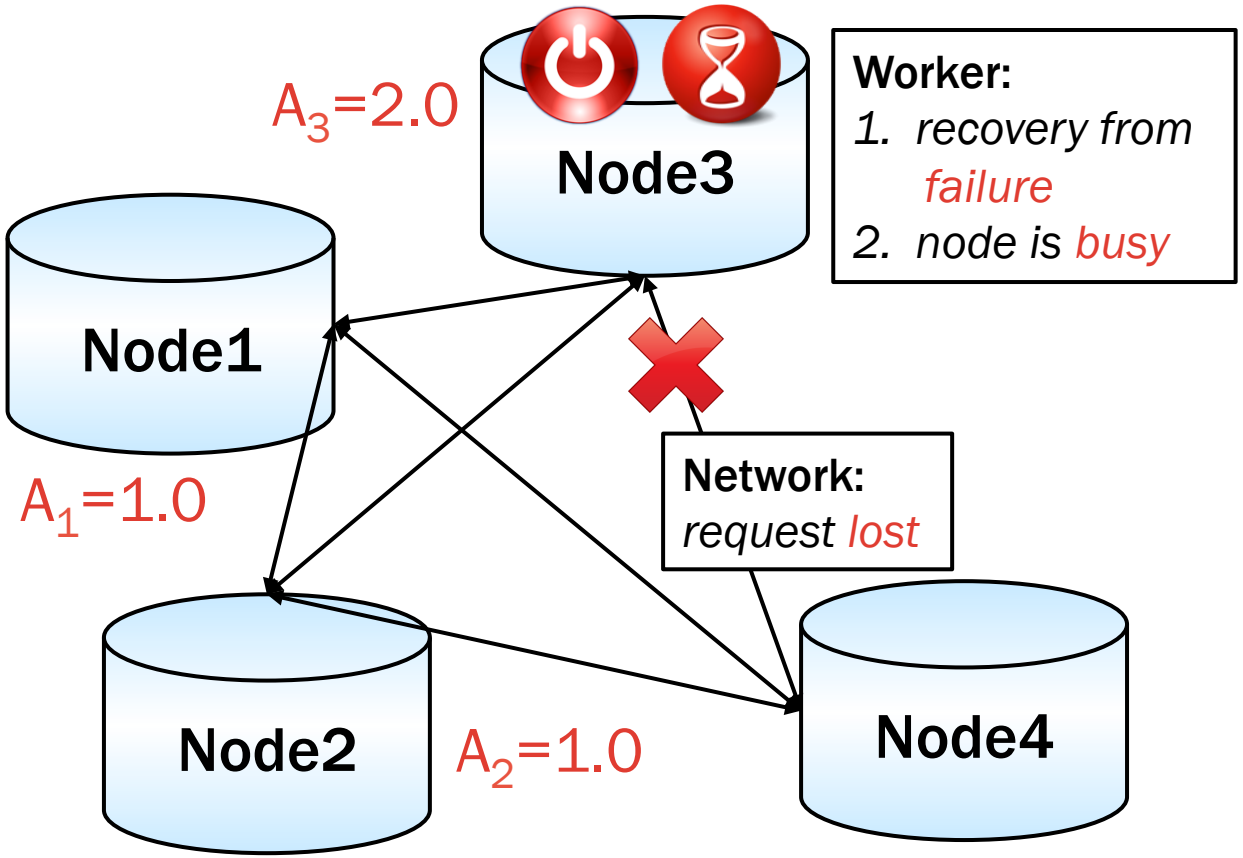


Data Replica in Cassandra



Replica Inconsistency

- **Different replicas of the same datum may have inconsistent values**



Replicas of A
@Node 1, 2, 3
A_1=1.0
A_2=1.0
A_3=2.0

Replica inconsistency
$A_1=A_2\neq A_3$

Why Replica Inconsistency?

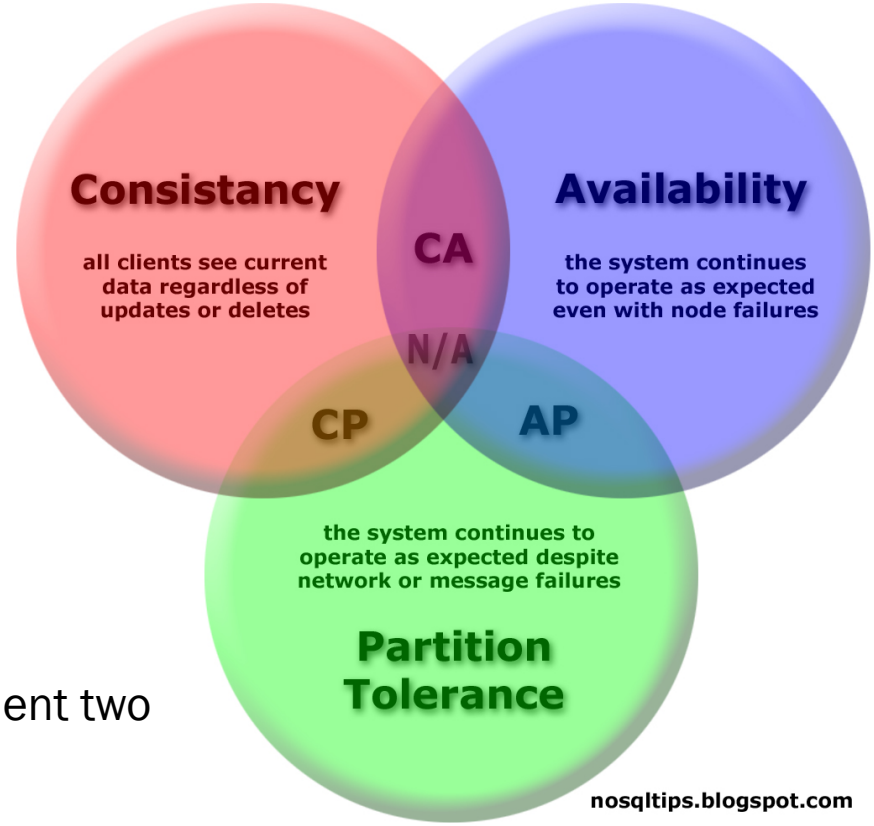
CAP theorem

Distributed data storage systems

cannot simultaneously satisfy

all these three features:

- ***Replica Consistency***
- ***System Availability***
- ***Network Partition Tolerance***

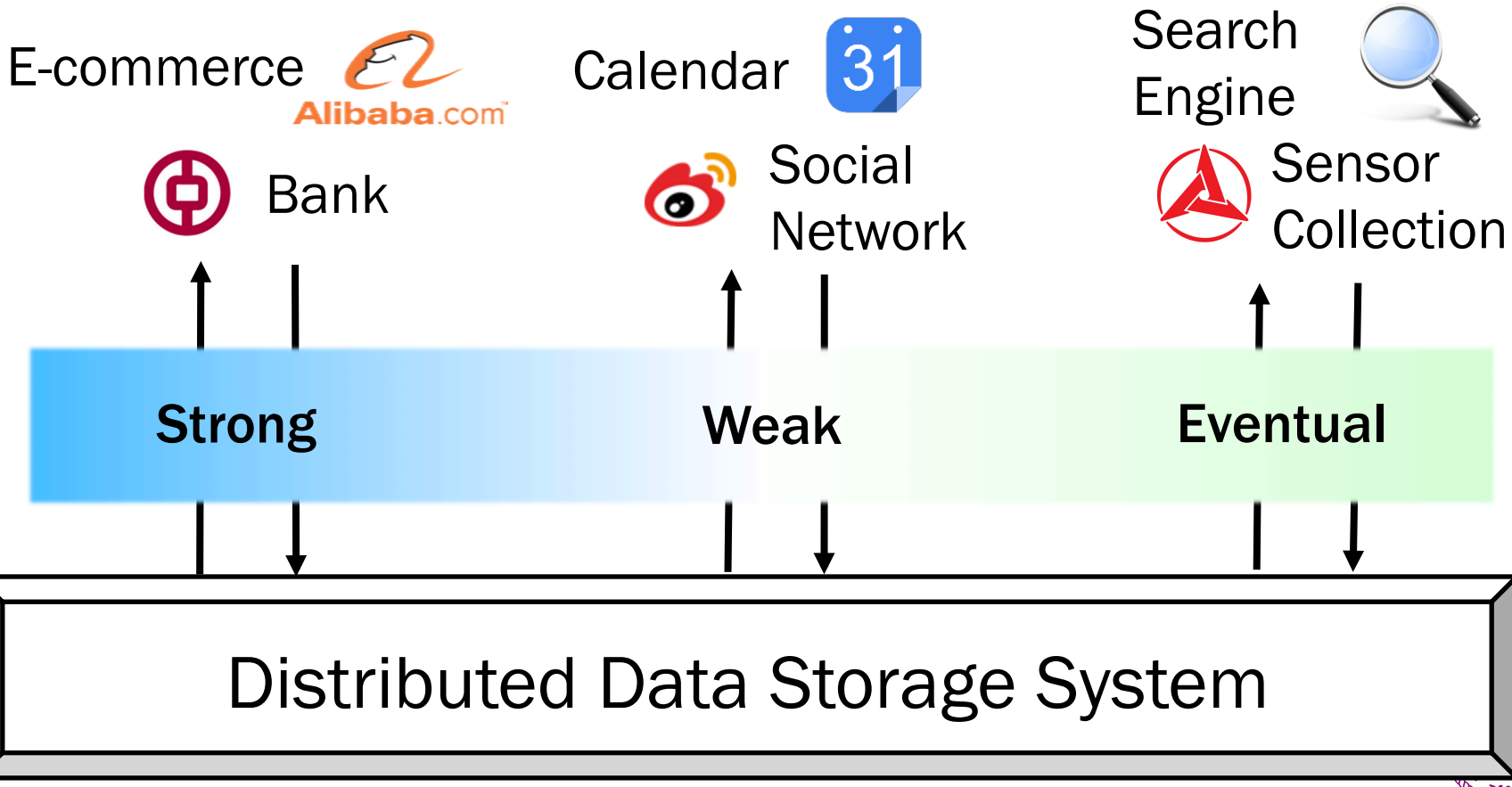


For best performance, most systems implement two features out of C/A/P:
Hadoop, Cassandra, MongoDB.....

S. Gilbert and N. Lynch. Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services. SIGACT News 02.

Consistency Requirements

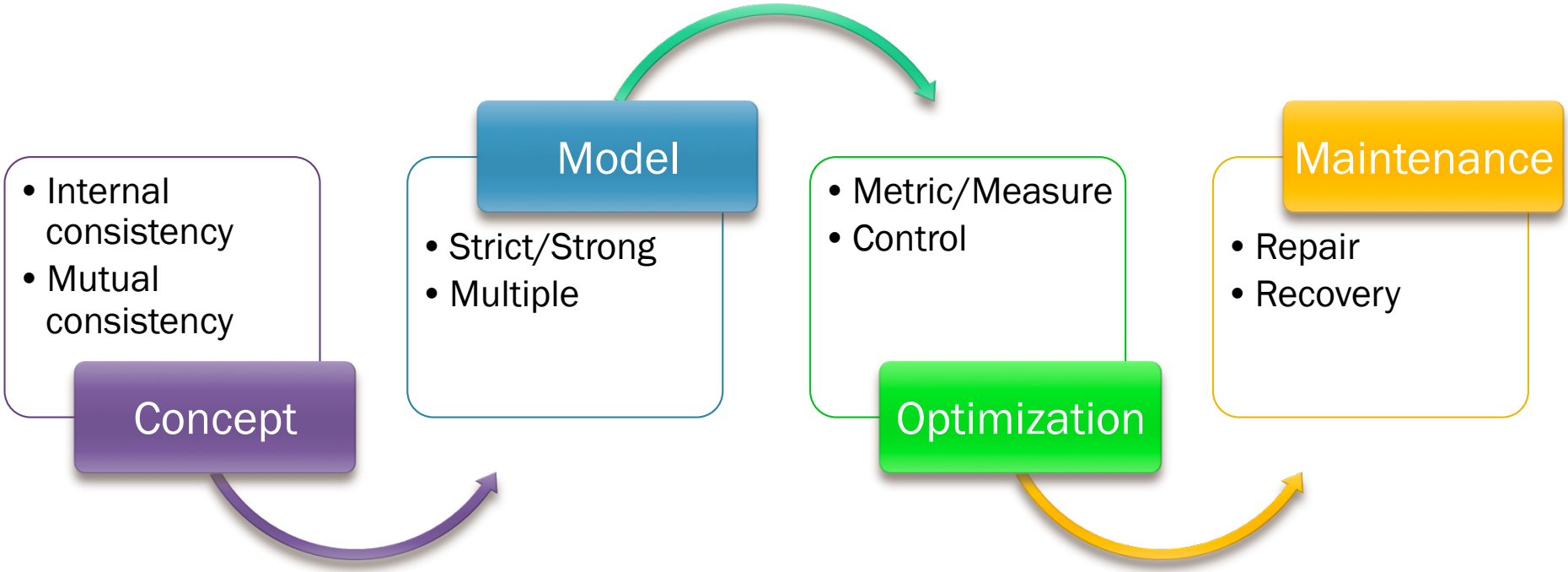
Replica consistency is complex but important for diverse big data applications



Overview

- Objectives
- Approach
- Progress
- Next Steps

Replica Consistency Issues



Replica Consistency Issues—Concept & Model

Concept
- **Internal consistency**
- **Mutual consistency**

Model
- **Strict/Strong**
- **Multiple**

Optimization
- Metric/Measure
- Control

Maintenance
- Repair
- Recovery

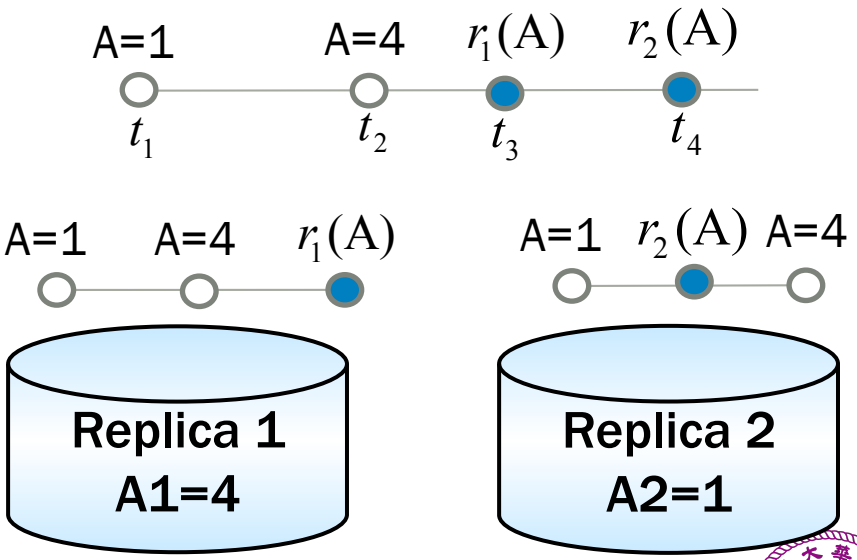
Consistency Model (Criteria)

Defines what executions of a distributed storage system are considered correct or a legal sequential history

Consistency
- Strong
- Weak

Architecture
- Data-centric
- Client-centric

Case of violating the Monotonous Read Model



Replica Consistency Issues—Optimization

- **Concept**
  - Internal consistency
  - Mutual consistency
- **Model**
  - Strict/Strong
  - Multiple
- **Optimization**
  - **Metric/Measure**
  - **Control**
- **Maintenance**
  - Repair
  - Recovery

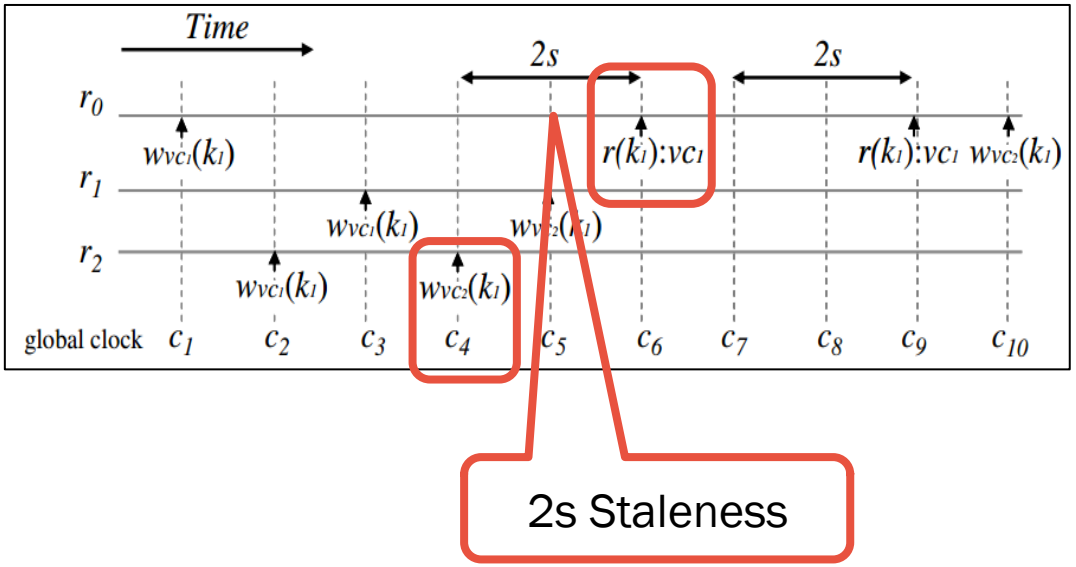
Consistency Metric

A metric indicating if system violates the consistency model

Consistency Measure

A quantization of the consistency violations

- UC Berkeley
  - PBS
  - Bolt-on Causal
- U Waterloo
  - DAX, Accordion
- HP Labs
  - DAG for Lamport's registers



Y. Zhu, P. S. Yu, and J. Wang. Latency Bounding by Trading off Consistency in NoSQL Store: A Staging and Stepwise Approach. 2012.

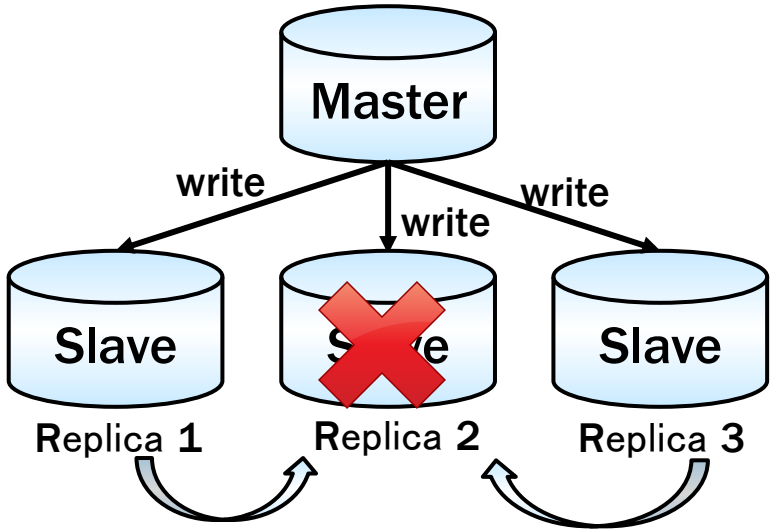
Replica Consistency Issues—Maintenance

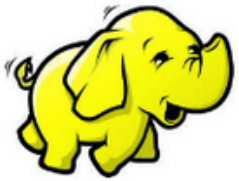
Concept
- Internal consistency
- Mutual consistency

Model
- Strict/Strong
- Multiple

Optimization
- Metric/Measure
- Control

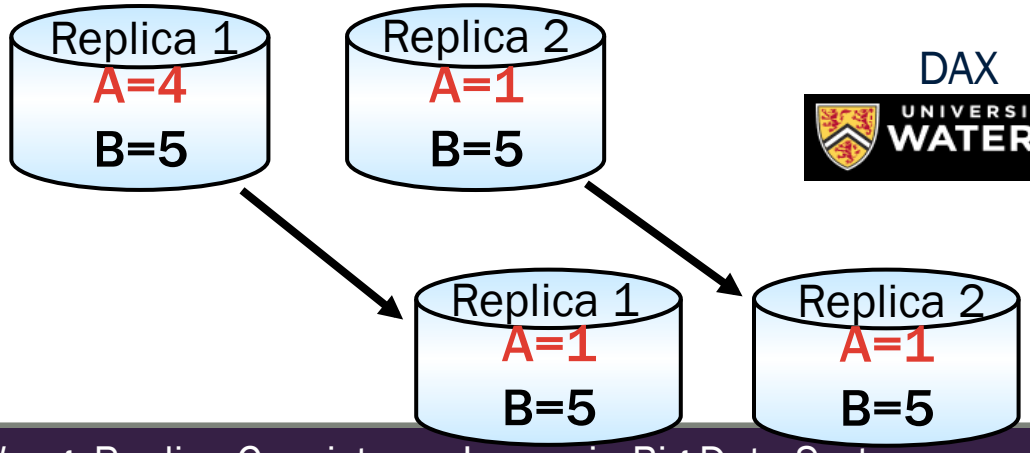
Maintenance
- **Repair**
- **Recovery**





mongoDB

riak

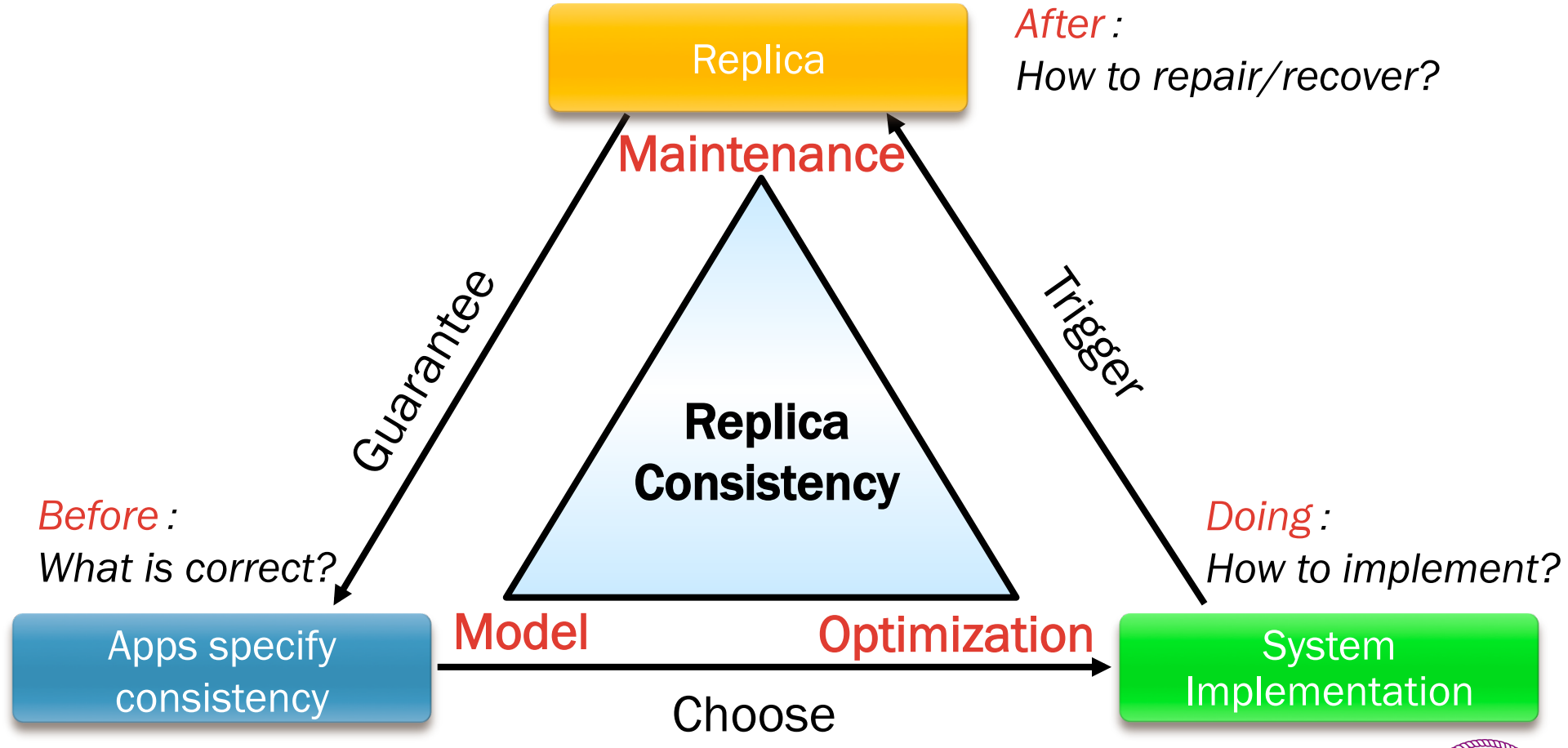


DAX

UNIVERSITY OF WATERLOO

cassandra

Replica Consistency Use Cases



Overview

- Objectives
- Approach
- Progress
- Next Steps

Consistency Measure

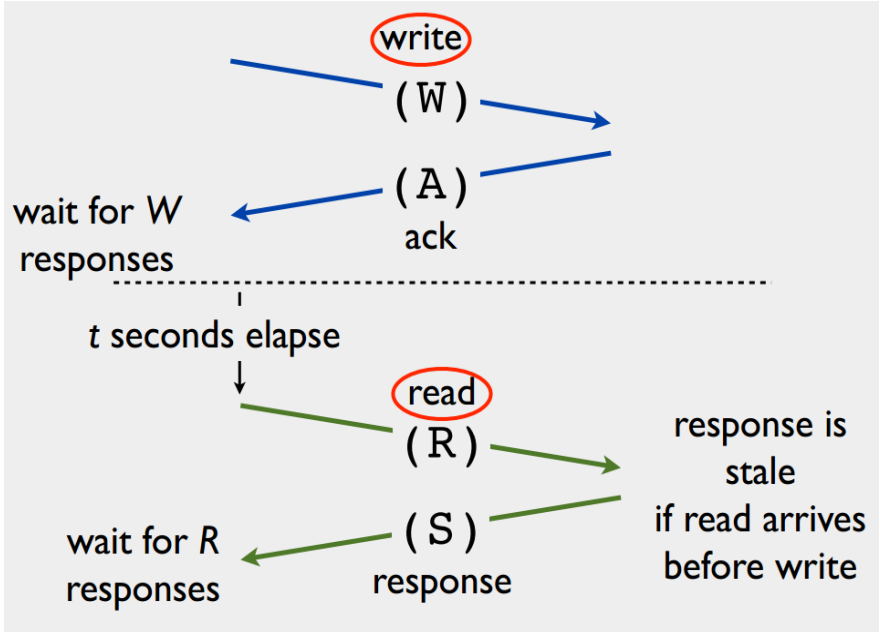
- **PBS: Probabilistically Bounded Staleness**
  - **A probabilistically bounded measure to quantify latency-consistency trade-offs**
  - **How eventual is eventual consistency? How consistent is eventual consistency?**

PBS

what: consistency prediction

why: weak consistency is fast

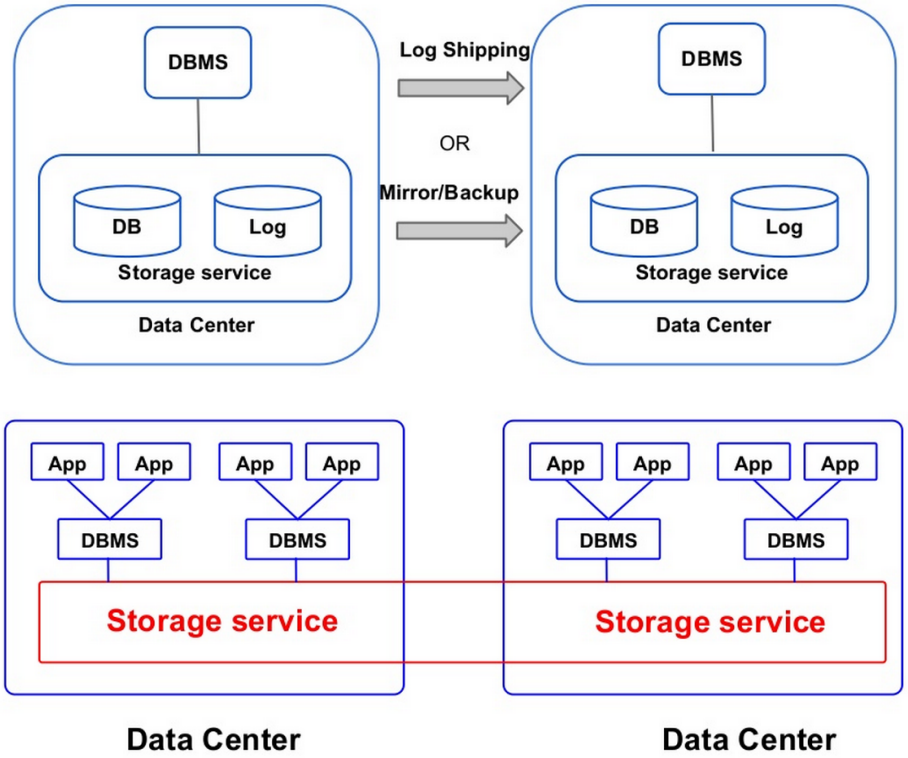
how: measure latencies
use *WARS* model



Peter Bailis, Shivaram Venkataraman, Michael J. Franklin, Joseph M. Hellerstein, Ion Stoica. Quantifying Eventual Consistency with PBS. PVLDB 2012.

Consistency Control

- **DAX: A Widely Distributed Multi-tenant Storage Service for DBMS Hosting**



Previous Solutions

- Complex
- Slow (synchronous)
- Or lost data (asynchronous)

DAX Solution

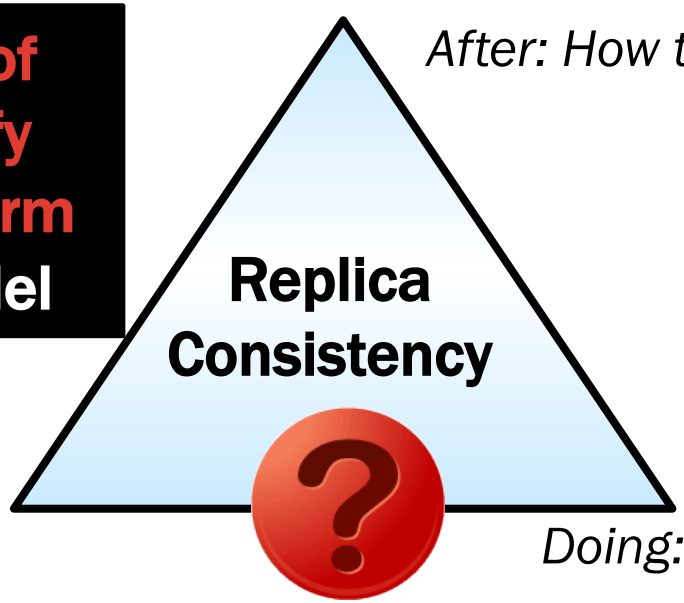
- Asynchronous Response to control the successful write operation
- DBMS Cache for versioning

Rui Liu, Ashraf Aboulnaga, and Kenneth Salem. DAX: A Widely Distributed Multi-tenant Storage Service for DBMS Hosting. PVLDB 2013.

Consistency Verification—Our Work@Tsinghua

- **Does the system conform its declared consistency model?**
  - **How to verify the system design and implementation?**
  - **How to analyze the causality underlying inconsistency?**

Develop implementation of consistency model to verify whether the system conform declared consistency model

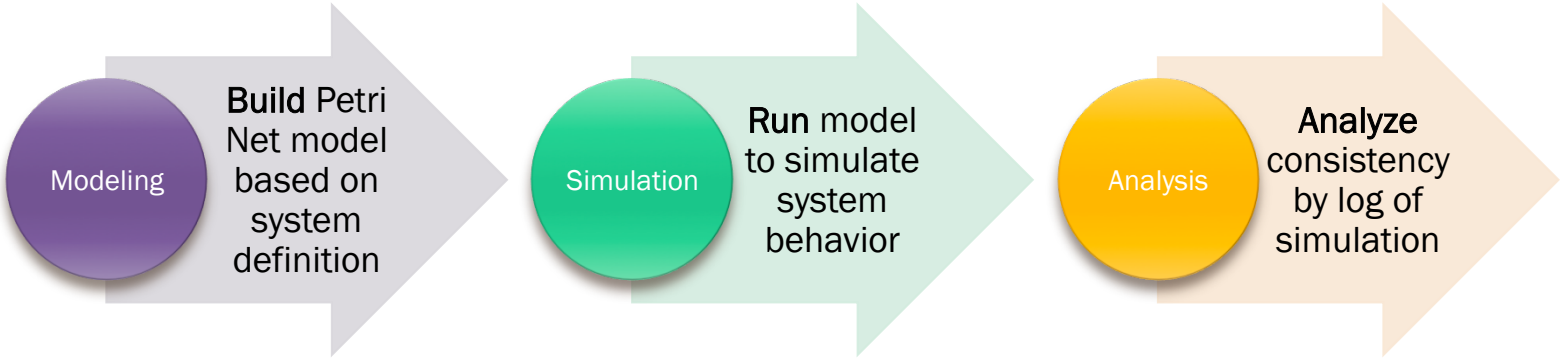


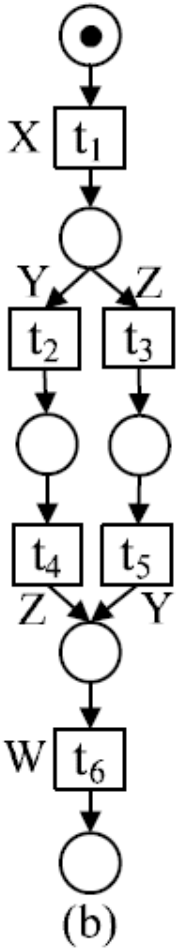
Case Studies

- Consistency Analysis with Petri Net in Cassandra
- Consistency Analysis with Queuing Theory in Cassandra

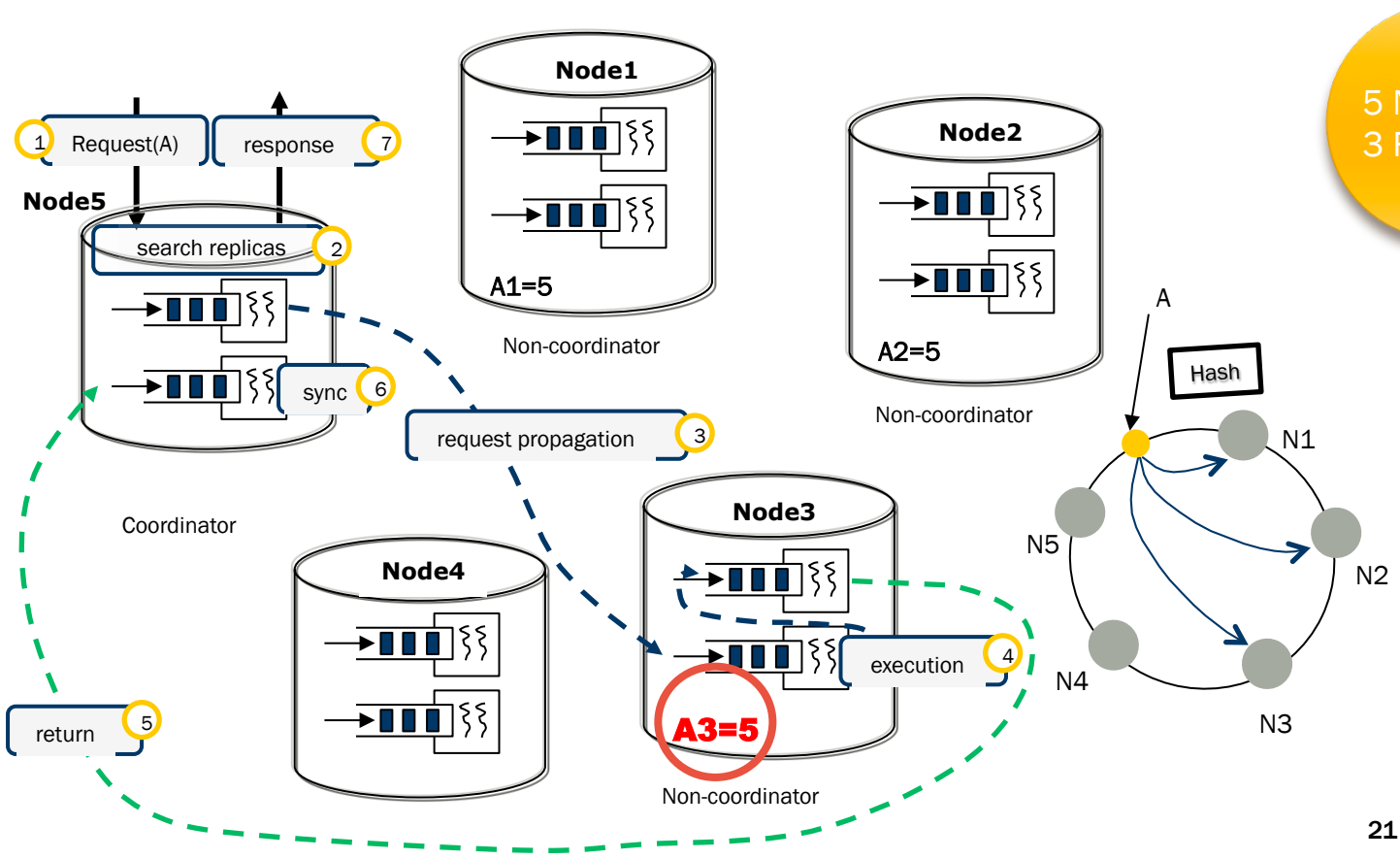
Petri Net Modeling Methodology

1. *Describe concurrency and synchronization*
2. *Simulation and analysis tools*
3. *Mathematical formal definition*

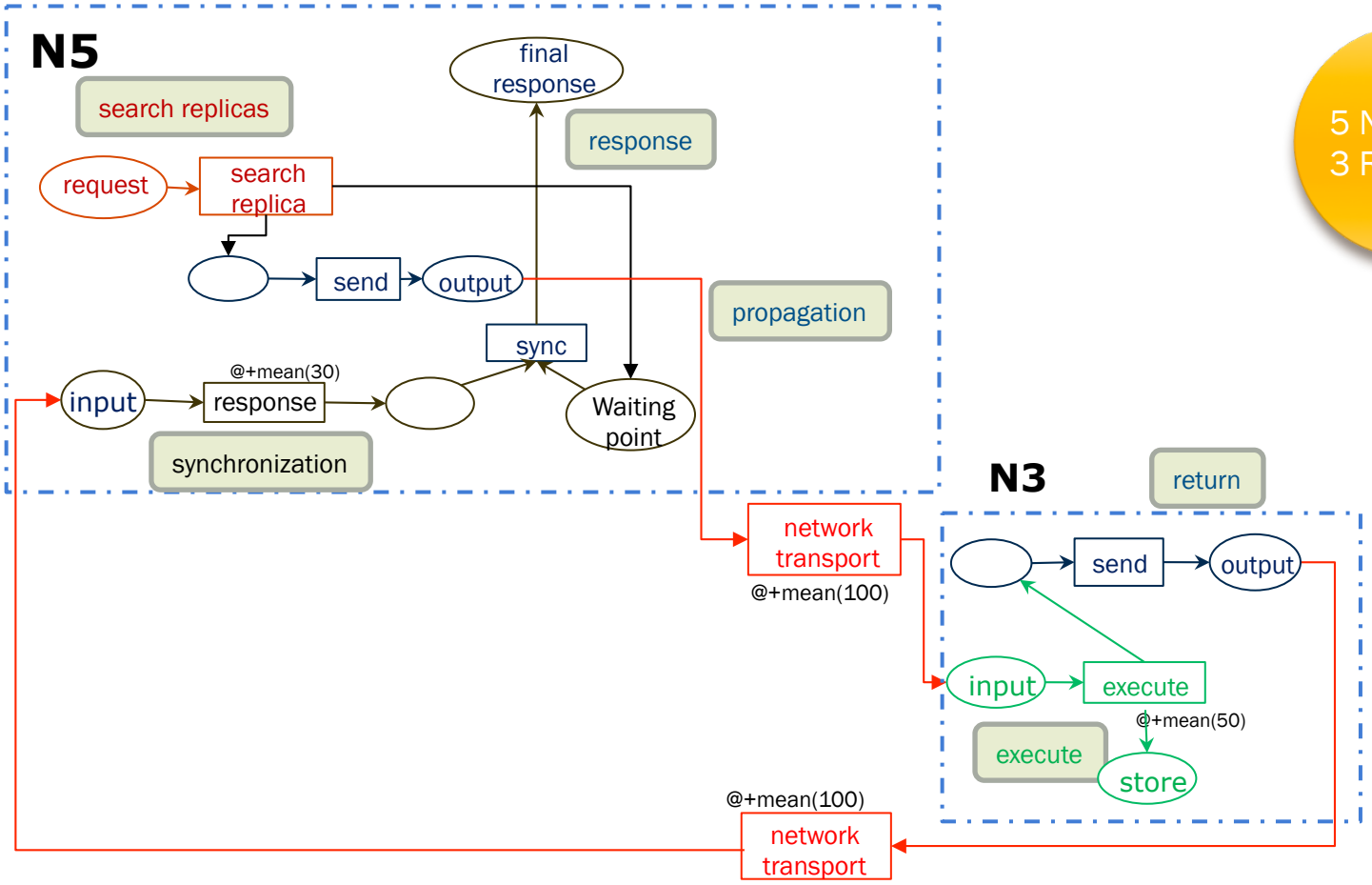




Write Process on Replica A3

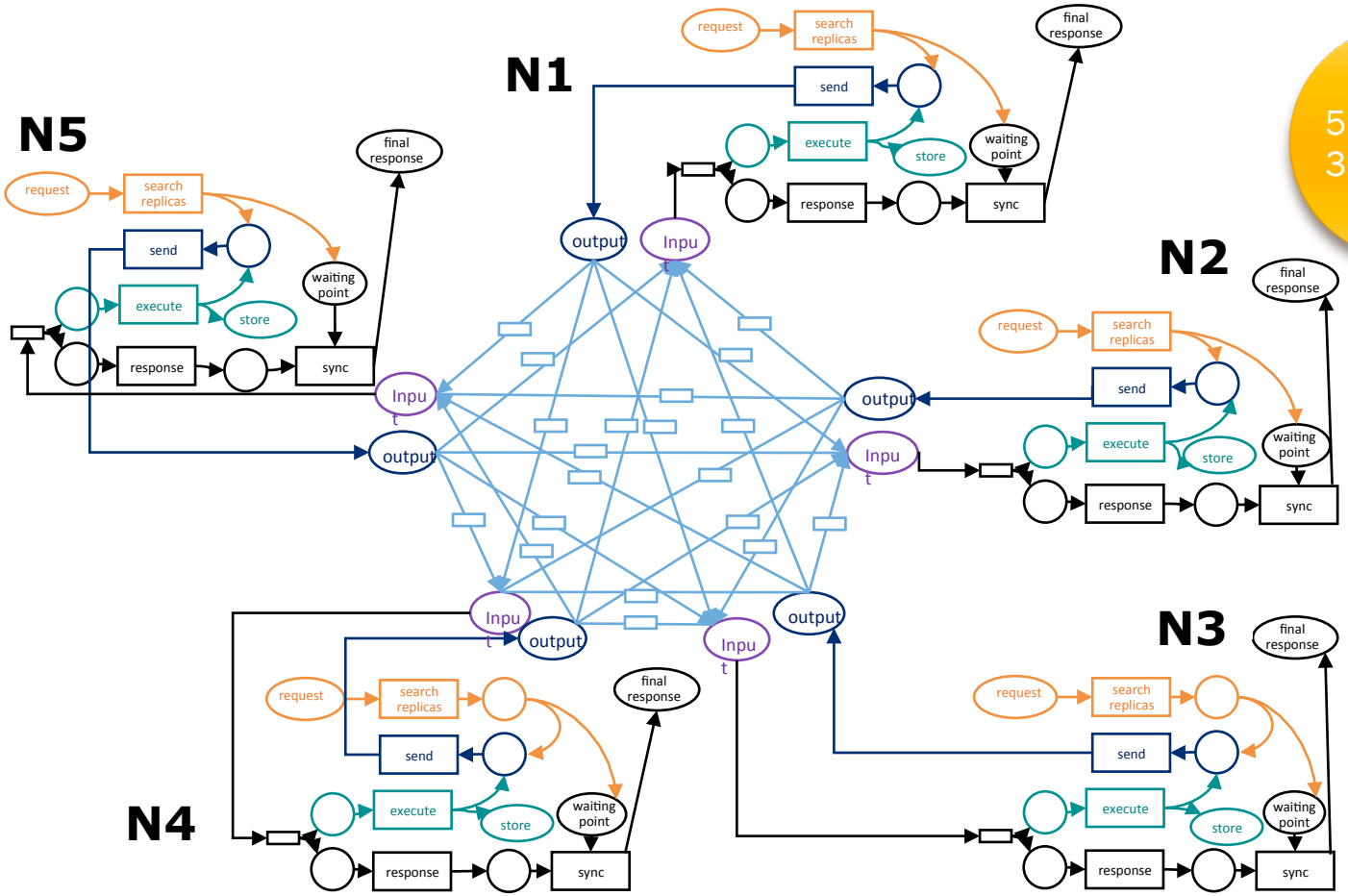


Partial Write Process

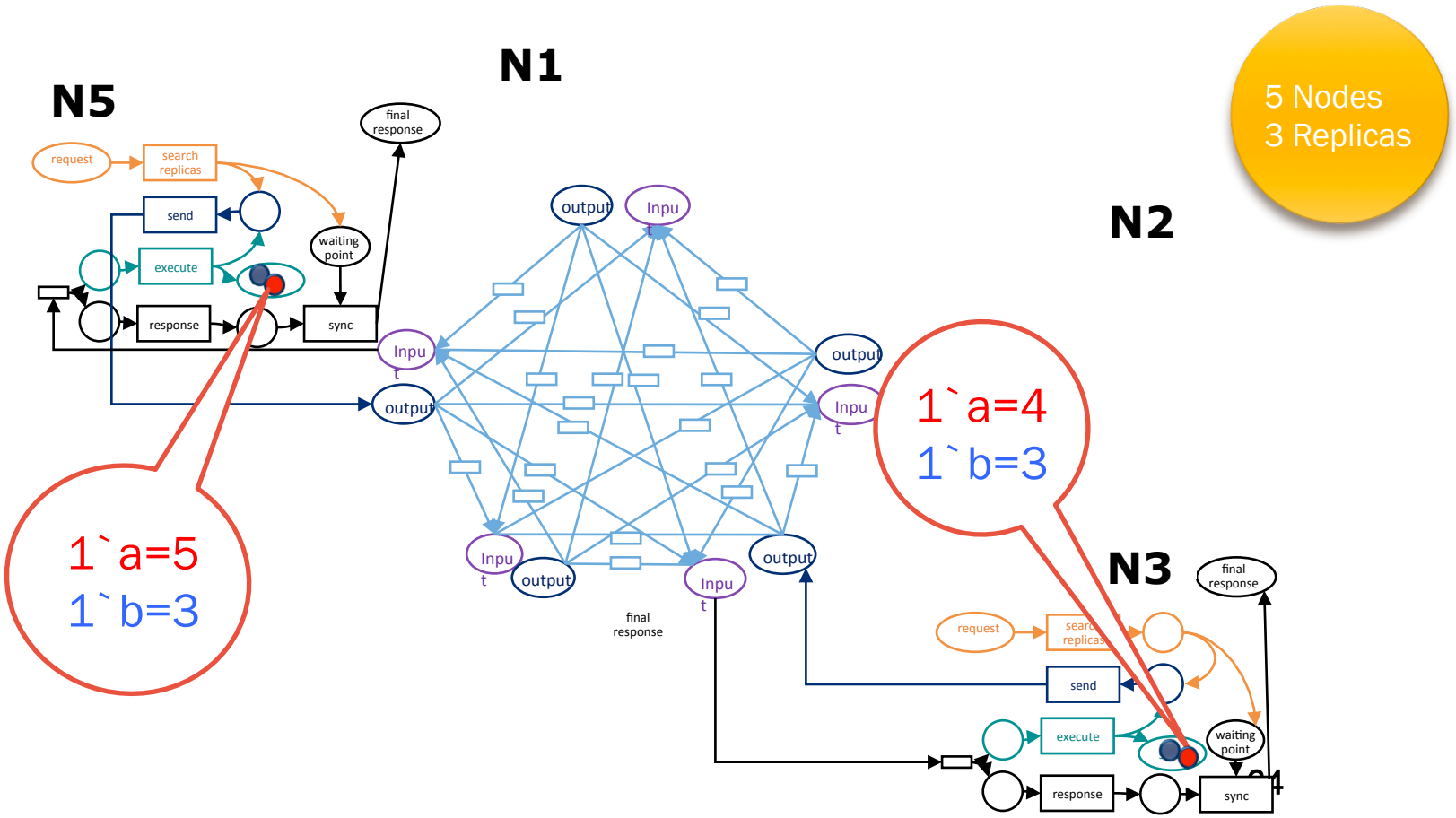


5 Nodes
3 Replicas

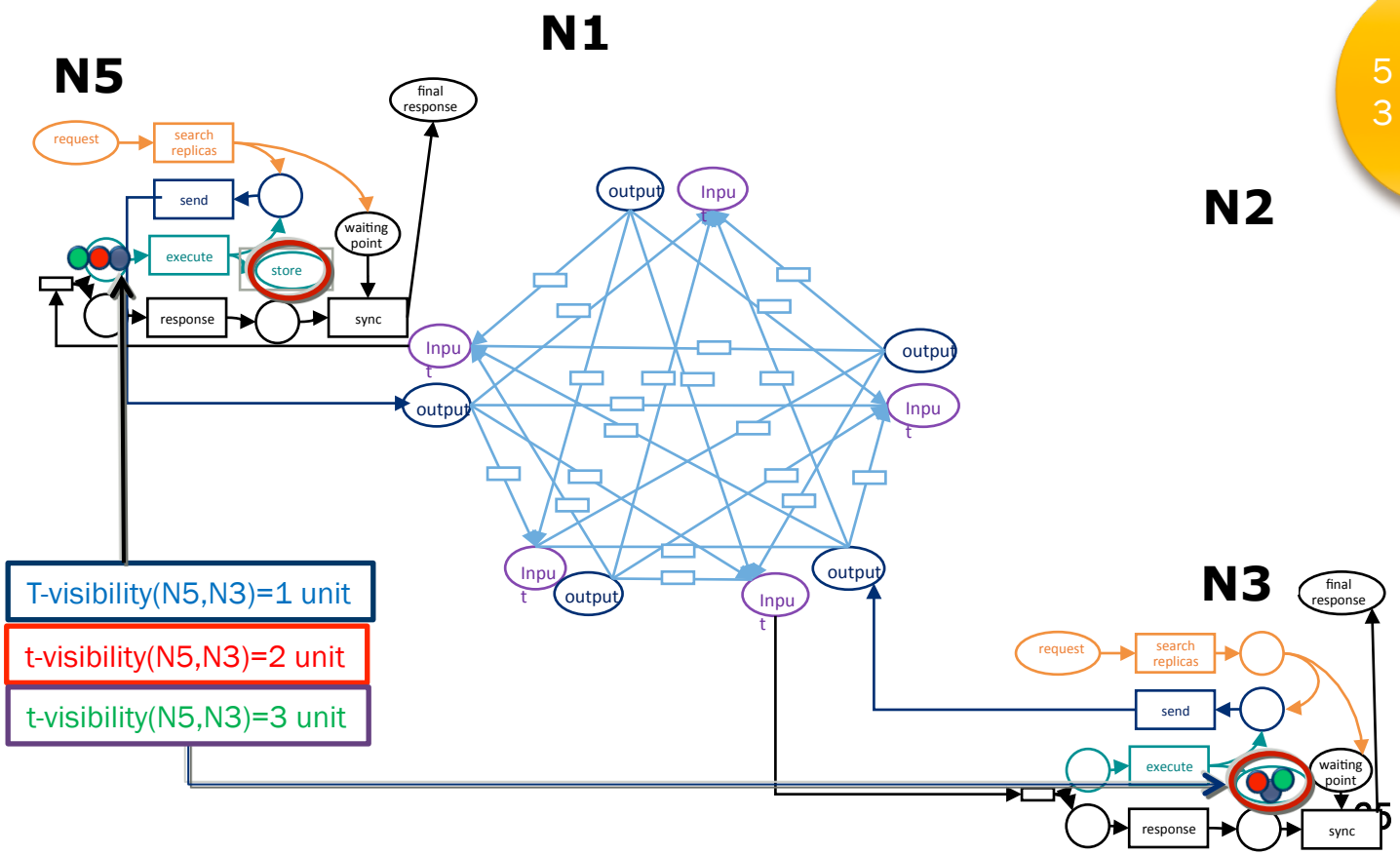
Whole Write Process Model



Inconsistency Detection



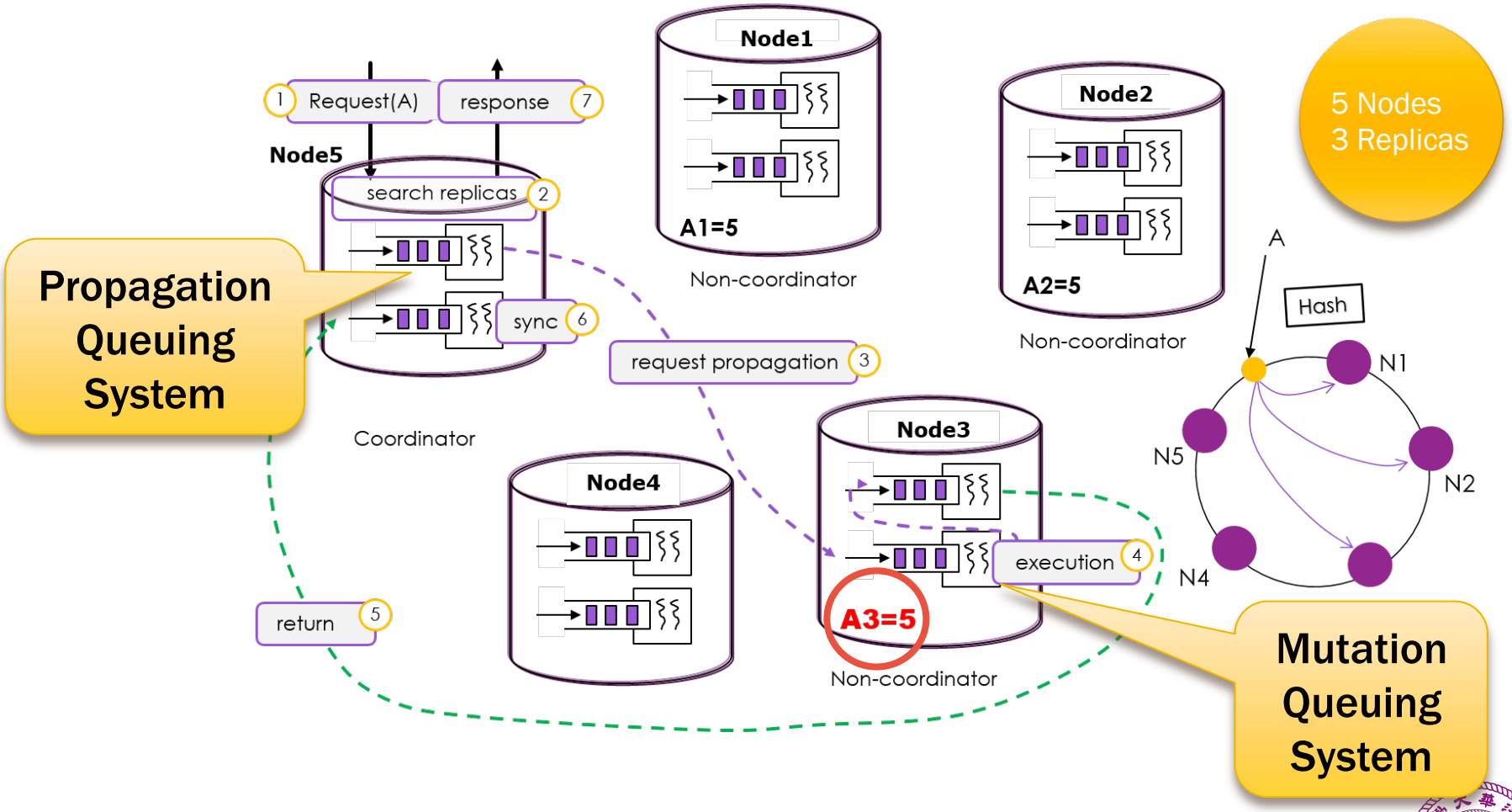
Inconsistency Measure



Case Studies

- **Consistency Analysis with Petri Net in Cassandra**
- **Consistency Analysis with Queuing Theory in Cassandra**

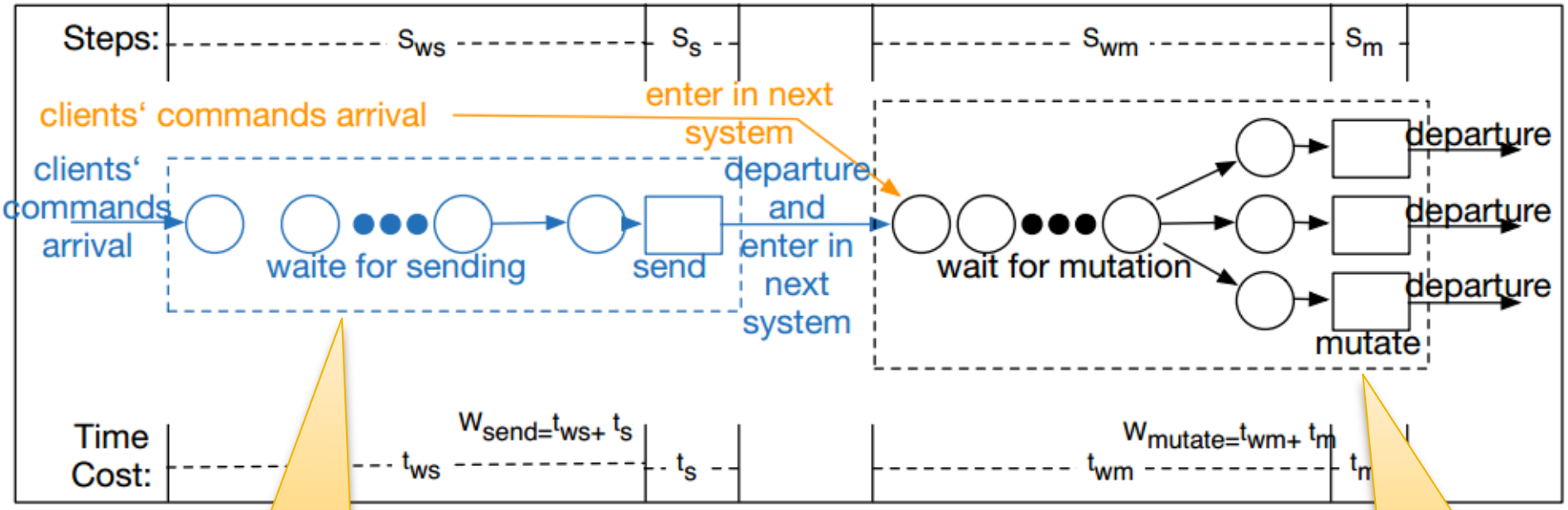
Write Process Modeling with Queuing Theory



Write Process Modeling with Queuing Theory

Replica consistency can be approximately computed by

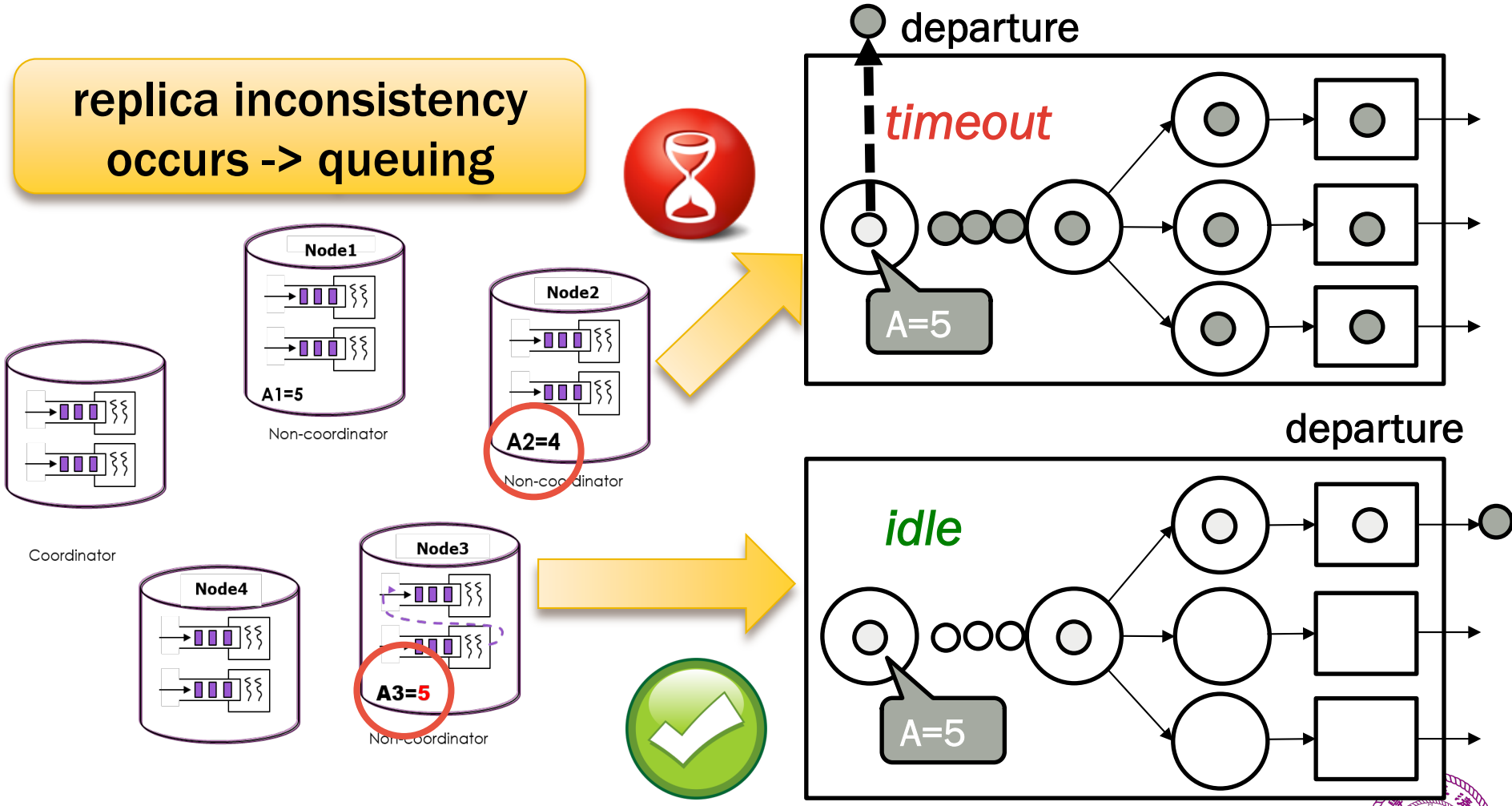
- **The queue strategies & the system services**



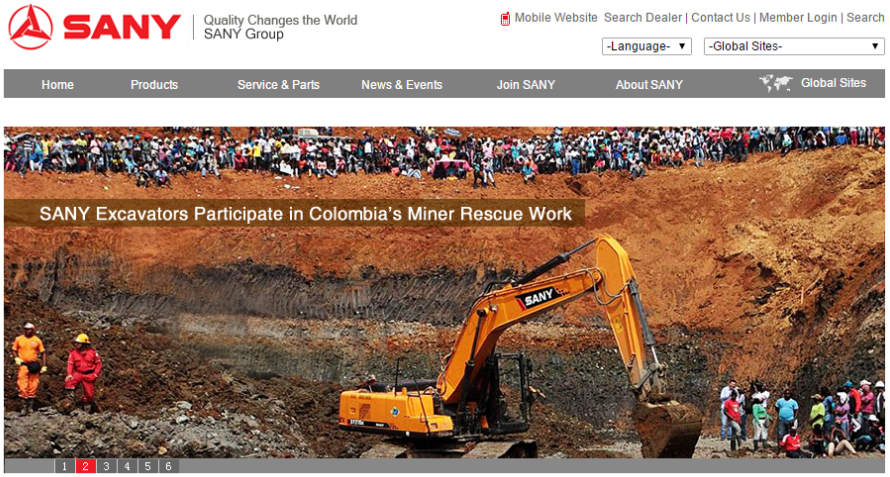
Propagation Queuing System

Mutation Queuing System

Consistency Analysis

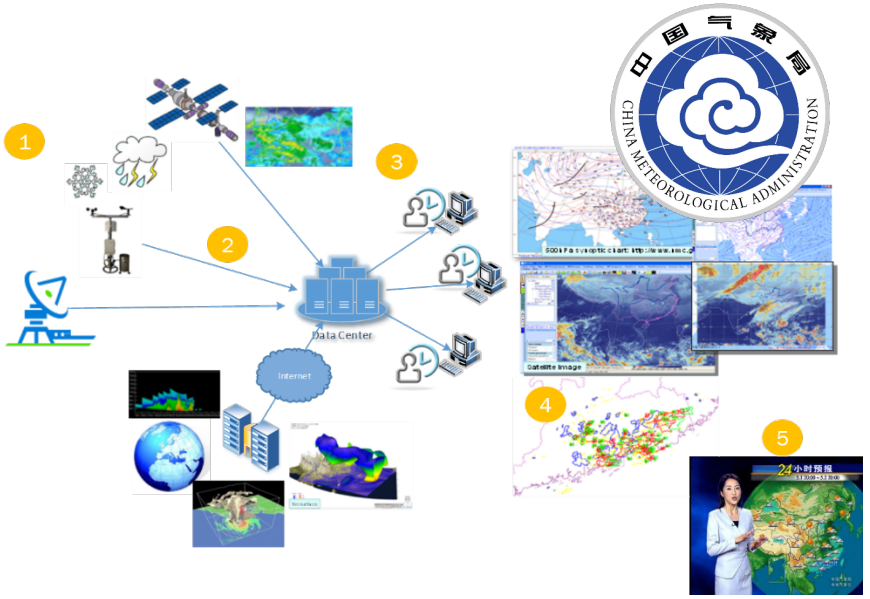


Industry Practices on Replica Consistency



Sensory Big Data

Weak Consistency



Meteorological Big Data

Weak & Strong Consistency

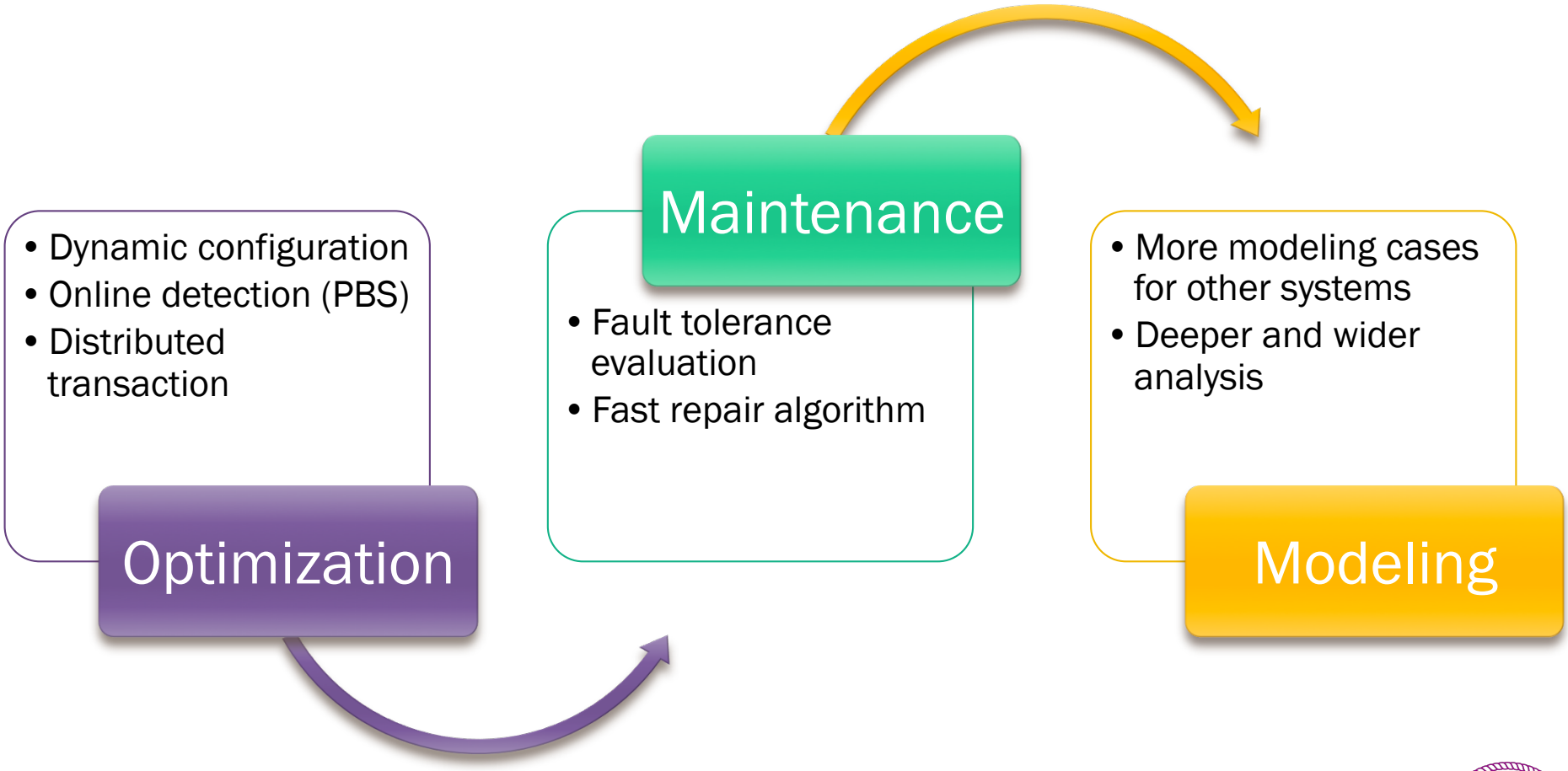
Healthcare Big Data

Weak & Strong Consistency

Overview

- Objectives
- Approach
- Progress
- Next Steps

Next Steps


Maintenance
• Fault tolerance evaluation
• Fast repair algorithm
Modeling
• More modeling cases for other systems
• Deeper and wider analysis
Optimization
• Dynamic configuration
• Online detection (PBS)
• Distributed transaction